Deep learning for understanding economic well-being in Africa from publicly available satellite imagery

Christopher Yeh[1], Anthony Perez[2], Anne Driscoll[3], George Azzari[2], Zhongyi Tang[4], David Lobell[3,5], Stefano Ermon[6], Marshall Burke[3,5]
[1]Department of Computing and Mathematical Sciences, California Institute of Technology
[2]Atlas AI
[3]Center on Food Security and the Environment, Stanford University
[4]Department of Economics, Boston University
[5]Department of Earth System Science, Stanford University
[6]Department of Computer Science, Stanford University
cyeh@caltech.edu, anthony@atlasai.co, anne.driscoll@stanford.edu,
george@atlasai.co, zztang@bu.edu, dlobell@stanford.edu,
ermon@cs.stanford.edu, mburke@stanford.edu

Abstract

Accurate and comprehensive measurements of economic well-being are fundamental inputs into both research and policy. Here we train deep convolutional neural networks to predict survey-based estimates of asset wealth across ~20,000 African villages from publicly-available multispectral satellite imagery with broad temporal and spatial coverage. Models are able to explain 70% of the variation in ground-measured village wealth in countries where the model was not trained, outperforming previous benchmarks from high-resolution imagery, and comparison with independent wealth measurements from censuses suggests that errors in satellite estimates are comparable to errors in existing ground data. Validating estimates of temporal changes in wealth across ~1,500 villages is also hampered by noise in training data, but district-aggregated satellite-based estimates explain up to 50% of the variation in ground-estimated changes in wealth over time, with daytime imagery particularly useful in this task. We quantitatively demonstrate the utility of satellite-based estimates for research and policy, and demonstrate their scalability by creating a wealth map for Africa's most populous country.

1 Introduction

Local level measurements of human well-being are important for informing public service delivery and policy choices, for targeting and evaluating livelihood programs by governmental and non-governmental organizations, and for the development and deployment of new products and services by the private sector. While recent work has generated granular estimates of a range of human and physical capital measures in parts of the developing world [6, 16, 26, 28, 31], similar data on key economic indicators remain lacking, constraining even basic efforts to characterize who and where the poor are.

For example, in the majority of African countries, at least four years pass between nationally-representative consumption or asset wealth surveys, the key source of data for internationally-comparable poverty measurements. Furthermore, these surveys have limited repeated observation of individual locations, complicating efforts to measure local changes in well-being over time. At current survey frequencies, each African household will appear in a household well-being survey on average less than once every 1,000 years, about two orders of magnitude less frequently than a household

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

in the U.S.A [33]. While not all households need to be observed to generate accurate economic estimates, sampling enough households to generate frequent and reliable national-level statistics is expensive, requiring an estimated $1 billion USD annual investment in lower-income countries to measure a range of indicators relevant to the Sustainable Development Goals [11]. Expanding these efforts to generate reliable local level estimates would add dramatically to these costs.

Although existing data are scarce and traditional survey methods are expensive to scale, other potentially relevant data for the measurement of well-being are collected with increasing frequency. For instance, while most African households are never observed in consumption or wealth surveys, their locations appear on average at least weekly in cloud-free imagery from multiple satellite-based sensors, and will have been observed in multispectral imagery at least annually for more than a decade. In our work, we show how such imagery can be used to accurately measure local-level well-being in Africa, including for countries where reliable survey data do not yet exist and where survey-based interpolation methods might struggle to generate accurate estimates.

We provide the following novel contributions: we train scalable models with high predictive performance (explaining at least 50% of village-level asset wealth variation) over 23 African countries using only publicly available imagery and data sources; to the best of our knowledge, we are the first to demonstrate the utility of daytime satellite imagery for estimating changes in economic well-being over time at the village and district levels; and we validate our estimates of economic well-being on practical economic policy and research tasks.

2 Related Works

Ground-based surveys remain the gold standard in understanding human well-being. In settings with limited survey data, Traditional approaches have used geostatistical models to predict health outcomes, standard of living, and housing quality in Africa [26–28, 31]. However, a growing body of research has applied machine learning techniques to estimate economic activity from non-traditional data sources. Earlier work demonstrated that coarse (1km/pixel) nighttime lights imagery can measure country-level economic performance over time [20], that convolutional neural networks (CNNs) on medium resolution daytime and radar backscatter satellite imagery can estimate granular population density [22], and that high-resolution ($<$1m/pixel) imagery from private-sector providers can be used to measure spatial variation in local economic outcomes in a handful of developing and middle-income countries [4, 10, 19, 23, 32]. Recent work has also demonstrated potential benefits of using object recognition methods on satellite imagery as a means of feature extraction for downstream prediction tasks [1]. Besides satellite imagery, other work has demonstrated the predictive power of mobile phone data [5] and natural language processing techniques on geotagged Wikipedia articles [30] for estimating economic well-being in Africa.

Our work is different from these existing works in several regards. First, we demonstrate performance improvements by combining nighttime lights imagery with daytime imagery as inputs to a single CNN model. Second, we demonstrate high predictive performance over 23 African countries, compared to 5 in Jean et al. [23] and Sheehan et al. [30]. Finally, we provide the first results on estimating changes in village and district level economic well-being over time from daytime satellite imagery.

3 Data

3.1 Survey and Census Data

We compiled data on asset wealth for $>$500k households living across 23 countries in Africa, drawn from 43 nationally-representative Demographic and Health Surveys (DHS) conducted between the years 2009 and 2016. We focus on asset wealth rather than other welfare measurements (e.g., consumption expenditure) as asset wealth is thought to be a less-noisy measure of households' longer-run economic well-being [14, 29], is a common component of "multi-dimensional" poverty measures used by development practitioners around the world, is actively used as a means to target social programs [2, 14], and is much more widely observed in publicly-available georeferenced African survey data. Following standard approaches [13, 29], for each household we compute a wealth index from the first principal component of survey responses to questions about ownership of specific assets. We pool all households in our sample in the principal components estimation such that the derived index is consistent over both space and time, and then average household values in

the cluster, the level at which geocoordinates are available in the survey data. (A "cluster" is roughly equivalent to a village in rural areas or a neighborhood in urban areas; in this paper we use the terms "cluster" and "village" interchangeably.) We removed clusters with invalid GPS coordinates and clusters for which we were unable to obtain satellite imagery, leaving us with 19,669 clusters. This approach assumes that assets contribute similarly to wealth across all countries in our data. Alternative methods of constructing the index using only directly observable subsets of these assets, or which allow the mapping of assets to the wealth index to differ by country and year, yield very similar wealth estimates, and the wealth index is highly correlated with log consumption expenditure (weighted $r^2 = 0.5$) in a small subset of countries where consumption data are available.

We also compiled asset wealth data for 9,000 households across over 1,400 clusters in 5 African countries from Living Standards Measurement Surveys (LSMS) bewteen 2005 and 2016. Unlike DHS data, the LSMS data form a panel—i.e., the same households are surveyed across multiple years—and thus enable the analysis of changes in household level economic well-being over time. Excluding households that were surveyed only once, we constructed a wealth index from the LSMS data using similar survey variables to the DHS data, also averaging to the cluster level, the lowest level with geocoordinates. While we cannot directly compare DHS and LSMS indices at the cluster level, district level estimates from the two sources are correlated ($r^2 = 0.60$). Furthermore, because LSMS data form a panel, we constructed an additional PCA-based index of changes in asset ownership that captures the principal component of wealth that is changing.

Finally, we gathered public census data from 8 countries that conducted a census within 4 years of a DHS survey in our main sample and whose surveys contained comparable asset variables to DHS. Our census sample contained 2,157,000 households observed in 656 administrative areas across these 8 countries. We aggregated the data to the second-level administrative boundaries provided with the census data (roughly corresponding to a "district"), the lowest level with georeferenced coordinates, using census-provided household weights to construct representative district averages.

3.2 Satellite Imagery

We obtained daytime Landsat surface reflectance and nighttime lights (NL) images centered on DHS and LSMS cluster locations from Google Earth Engine [15]. We used 3-year median composite images captured by the Landsat 5, Landsat 7, and Landsat 8 satellites over four periods: 2005-08, 2009-11, 2012-14, and 2015-17. Each composite is created by taking the median of each cloud free pixel available during that period. The motivation for using three-year composites was two-fold. First, multi-year median compositing has seen success in similar applications as a method to gather clear satellite imagery [3], and even in 1-year compositing we continued to note the substantial influence of clouds in some regions, given imperfections in the cloud mask. Second, the outcome we are trying to predict (wealth) tends to evolve slowly over time, and we similarly wanted our inputs to not be distorted by seasonal or short-run variation. The images have a spatial resolution of 30m/pixel with 7 multispectral (MS) bands: red, green, blue, near infrared, 2 shortwave infrared bands, and thermal.

We also created 3-year median composites for NL imagery. Because no single satellite captured nightlights for all of 2009 to 2016, we used DMSP [21] for the 2005-08 and 2009-11 composites, and VIIRS [9] for the 2012-14 and 2015-17 composites. DMSP and VIIRS images have incompatible units and different resolutions, so we treat them as separate image bands in our models. The images are resized using nearest-neighbor upsampling to cover the same spatial area as the Landsat images.

Both MS and NL images were cropped to dimension 224×224, the input size of our CNN architecture, spanning 6.72 km on each side (30m Landsat pixel size $\times$ 224px = 6.72km). Thus any survey cluster whose location coordinates are artificially displaced by more than 4.75 km ($6.72/\sqrt{2}$) is completely beyond the spatial extent of the satellite imagery. Each band is normalized to have mean 0 and standard deviation 1 across our entire dataset.

4 Methods

We trained CNN models based on the ResNet-18 (v2) architecture [18] to predict the DHS and LSMS cluster-level wealth indices from satellite imagery. The CNNs are trained for 150 epochs (200 epochs for DHS "out-of-country") with an Adam optimizer [24], mean squared-error loss function, and batch size of 64. We performed grid search over learning rate and L_2 weight regularization hyperparameters.

In addition to training separate models on MS bands and NL bands, we also tested a "combined" model which concatenates the final layers of the MS and NL models and has a ridge-regression model on top. We found that this approach performed better than stacking the nightlights and Landsat bands together in a single model. Implementation details can be found in our code[1].

For DHS data, an average of 25.59 households (standard deviation = 5.59) were surveyed for each cluster, compared to an average of 6.37 households (sd = 3.57) in LSMS. Due to the lower number of households surveyed for LSMS, which results in noisier estimates of cluster-level wealth, we weighted LSMS clusters proportional to their surveyed household count in the loss function during training. We did not weight DHS clusters.

To evaluate how models perform in even more data-limited situations, we trained our deep models on random subsets of 5%, 10%, 25%, 50%, and 100% of the full training data, and we report the mean r^2 over 3 repeated trials with different random subsets (Fig. 1c).

Transfer Learning Models We compared our end-to-end training procedure with a transfer learning approach inspired by Jean et al. [23]. In this approach, we take nightlights as a noisy but globally-available proxy for economic activity ($r^2 \approx 0.3$ with DHS wealth index), and a model is trained to predict nighttime lights values from daytime multispectral imagery. This process summarizes high-dimensional input daytime satellite images as lower-dimensional feature vectors that are then used in a regularized regression to predict wealth. However, we note that our transfer learning experiments are not directly comparable to results in Jean et al. [23] due to two differences. First, we treat nightlight prediction as regression instead of 3-bin classification. Second, our experiments span 23 countries, much larger than the 5 countries than Jean et al. [23] focus on.

Because our images have a mixture of DMSP and VIIRS values, and the two satellites have different spatial resolutions, we framed transfer learning as a multitask regression problem. We extracted the neural network's final layer output predictions for both the DMSP value and the VIIRS value, and regressed on whichever nightlights label was available for each daytime image. On the nightlights prediction task over locations sampled from all 23 DMSP countries, our transfer learning models achieved performance of $r^2 = 0.82$ when using RGB bands and $r^2 = 0.90$ when using all Landsat bands. With these models trained to predict nightlights values from daytime imagery, we froze the model weights and fine-tuned the final fully-connected layer to predict the wealth index.

Baseline models We train simpler k-nearest neighbor models ("KNN") on nightlights that predict wealth in a given location i as the average wealth over the k locations with nightlights values closest to that in i, where k is tuned by cross-validation. The KNN model allows a non-linear and non-monotonic mapping of nightlights to wealth. We also train a regularized linear regression on scalar nighlights ("scalar NL") as a baseline model.

Data Splits and Cross Validation In all of our experiments, we used 5-fold cross-validation. For the DHS "out-of-country" tests, we manually split the 23 countries into 5 folds such that each fold had roughly the same number of villages. For DHS "in-country training," we split the 19,699 villages into 5 folds such that there was no overlap in satellite images of the villages between any fold, where "overlap" is defined as any area (however small) that is present in both images. We used the DBSCAN algorithm to group together villages with overlapping satellite images, sorted the groups by the number of villages per group in decreasing order, then greedily assigned each group to the fold with the fewest villages. We followed the same procedure to create 5 LSMS "in-country" folds. We did not perform "out-of-country" tests with LSMS data.

For each of the input band combinations (MS, NL, MS+NL), we trained 5 separate models, each with a different "test" fold. Of the four remaining folds, three folds were used to train the models, with the final fold serving as the validation set used for early stopping and tuning hyperparameters. We report our results on the five test folds grouped together.

5 Results

Predictive performance over space Our combined model is predictive of cluster-level asset wealth, with predictions explaining on average 70% of the variation in ground-based wealth measurements

[1] `https://github.com/sustainlab-group/africa_poverty`

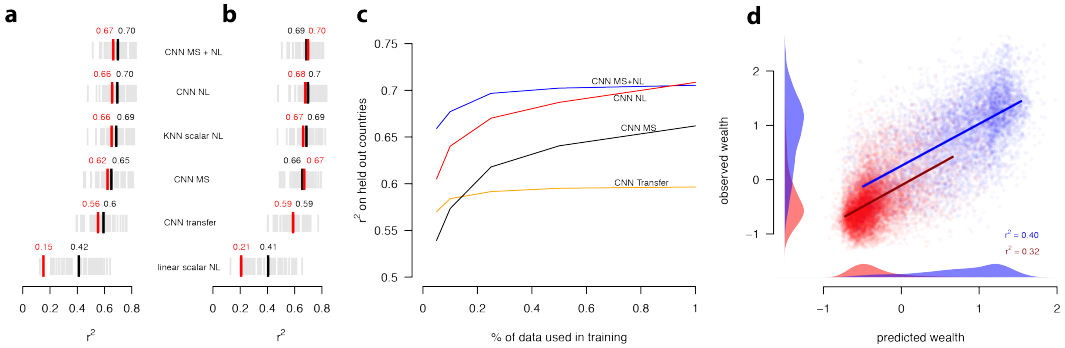

Figure 1: **Performance by model and across different samples**. **a** Predictive performance of satellite predictions trained using 5 different machine learning models. `CNN transfer` = transfer learning on nightlights with RGB Landsat imagery. Each grey line indicates the performance (r^2) on a held-out country-year, black lines and text show the average across country-years, and red lines and text show the r^2 on the pooled sample. **b** As in (a) but for evaluation on held-out villages within the same country. **c** Performance by amount of training data used. **d** Performance of `CNN MS + NL` model in urban versus rural regions in held-out countries. Model is trained on all data in training set and then applied separately to either urban or rural clusters in held-out countries. Each dot is an urban (blue) or rural (red) cluster.

in held-out country-years (Fig. 1a). Performance in individual held-out countries is never below 50% of variation explained, and often exceeds 80% (median = 70.4%, Fig. 1a,b), indicating our model is not simply separating wealthier African countries from poorer countries, but capably differentiating wealth levels within countries. These results exceed performance in earlier work on a similar task using high-resolution imagery [19, 23] or mobile phone data [5] as input, and match or exceed benchmarks for in-country performance from geostatistical models used to predict health outcomes, standard of living, and housing quality in Africa [26–28, 31]. Visualization of model activation maps suggests that the model learns semantically-meaningful features that are intuitively related to wealth, including filters for urban areas, agricultural regions, water bodies, and deserts. Aggregating predictions and ground measurements to the district level further improves performance, with predictions explaining on average 83% of the ground measurements in held-out countries not used to train the model. Improved performance with aggregation is consistent with errors cancelling when either the predictions or ground data are averaged.

Notably, CNNs trained on only MS or only NL imagery perform similarly to each other and almost as well as the combined model (MS+NL), suggesting that these two inputs contain similar information for the task of predicting spatial variation in African wealth (Fig. 1a-b). Consequently, our approach of directly using nightlight images as model inputs performs better than using them indirectly as a proxy, as in an earlier transfer learning approach [23]. This trend remarkably holds for highly data-limited settings: even when trained on data from only 5% of the surveyed clusters ($n < 1000$), our best models trained end-to-end outperform transfer learning (Fig. 1c).

The KNN model that predicts wealth in a given location from wealth in locations with similar nightlights values performs nearly as well as the deep learning models in predicting spatial variation, and much better than a linear model using scalar nightlights as input (Fig. 1a-b) – although neither nightlights models are predictive of temporal changes in wealth, while daytime models are. These results suggest that non-linearities and spatial structure in nightlights is important for explaining spatial variation in wealth, and may explain why the transfer learning approach, which only predicts a scalar nighttime light intensity from daytime images, performs worse than end-to-end training.

Predictive performance over time Many research and policy applications require estimates of changes in economic measures over time as well as over space. There are important challenges, however, in using available ground surveys to measure changes in economic outcomes over time at a local level, as well as in evaluating our deep learning approach's ability to do so. First, most existing surveys do not repeatedly measure outcomes at the same locations over time, i.e. they are not "panel" data; the DHS surveys, for instance, draw a new sample of clusters each survey round.

Second, temporal changes over a few-year time span are likely to be small relative to cross-sectional differences, and any random noise in each year's survey will diminish the signal in these changes.

Given these challenges, we take three approaches to measuring and predicting changes in wealth over time. We first use repeated rounds of DHS surveys and spatially match a cluster in one survey year to the nearest cluster in a previous survey year (ignoring the random noise added to the village locations by the survey team; see below), and compute wealth changes as the difference in wealth index between matched pairs of clusters. Second, we use an independent smaller set of houeshold level panel data from LSMS to construct cluster-level changes in a wealth index. In both cases, predictions from a deep learning model using imagery as input can explain between 15-17% of the variation in survey-measured changes in asset wealth in held-out villages. In contrast to our cross-sectional results, deep learning models using nightlights as input performed significantly worse than models using multispectral imagery ($r^2 = 0.15$ vs. $r^2 < 0.01$), likely because nightlights show little variation over time in our sample locations. While temporal performance in multispectral models remains low relative to our model's performance in cross section, exceeding this temporal performance would be difficult for any model as the small average temporal change in wealth in our sample (0.08 standard deviations of our wealth index) could easily be obscured by noise in the two survey values being differenced.

In a third experiment, we use the same LSMS data to construct a PCA-based index of changes in asset ownership (rather than a change in indexes, as before) to better capture the component of wealth that is actually changing. By construction this index has greater variation over time, and satellite-based features are more predictive ($r^2 = 0.35$), again with the models requiring multispectral imagery inputs in order to perform well (Fig. 2a). As in the cross-sectional results, models again appear to learn features related to urbanization and to changing agricultural patterns. Aggregating ground- and satellite-based estimates to the district level again leads to substantial performance improvements (Fig. 2b), with predictions of asset wealth changes explaining up to 50% of the ground-estimated changes in asset wealth. Improved performance with aggregation is again consistent with errors cancelling when either the predictions or ground data are averaged. To our knowledge, these are the first known remote-sensing based estimates of local-level changes in economic outcomes over time across a broad developing country geography, and provide benchmarks for future work.

Understanding model performance While some of the combined model's overall performance in spatial prediction derives from distinguishing wealthier urban areas from poorer rural areas, the model is still able to distinguish variation in wealth within either rural or urban areas (Fig. 1d). In either case, much of the model's explanatory power, at least in cross section, appears to be in separating wealthier clusters from poorer clusters rather than in separating the poor from the near poor (Fig. 1d). Performance at the country level (as shown in Fig. 1a) is not strongly related to country-level statistics on headcount poverty rates, urbanization, agriculture, or income inequality, although we do find that model performance is somewhat worse in settings where within-village variation in wealth is high. Poorer performance in these settings could be because our model has difficulty making accurate predictions in locally heterogeneous environments (a problem likely amplified by the random noise that has been added to the data; see below), or because sample-based estimates from the ground surveys are themselves likely noisier when local variation is high.

Other sources of noise in the ground data (e.g., due to survey recall bias, sampling variation or geographic inaccuracies) could also worsen model performance. To explore the overall role of ground-based error in model performance, we take two approaches. First, we compare both model-based and ground-based measures against an independent measure of asset wealth derived from census data in eight countries, with the comparison made at the district level, the lowest level of geographic identification available in public census data. We find that ground-based measures are only slightly more correlated with this independent wealth measure than our model-based estimates (weighted r^2 averaged across country-years: 0.89 vs. 0.83). This suggests that at least some of the prediction error in our main results derives from noise in the survey data.

Second, a known source of error in our ground data is random noise added to cluster-level geo-coordinates by survey implementers to protect privacy. In practice this "jitter" creates geographic misalignment between our input imagery and the true location of the surveyed villages. To understand the performance cost of this noise, we iteratively add additional locational noise to our training data and then re-evaluate model performance on test data which are either also additionally jittered or not. Performance degrades with additional jitter, although much less rapidly when evaluating on data that

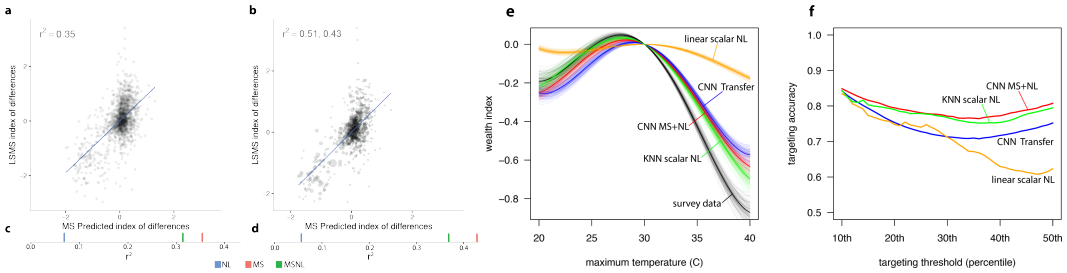

Figure 2: **Satellite predictions of ground-measured changes in wealth over time. a** Performance of satellite-based model trained to predict the index of the change in wealth over time. **b** Same as (a), but with observations aggregated to the district level. Dot size represents number of village observations in each district, and r^2 is reported both weighted and unweighted by number of villages. **c-d** Cross-validated r^2 of models trained on multispectral (MS, red), nightlights (NL, blue), and both (MSNL, green). Every reported r^2 in (c) and (d) is unweighted.

Using satellite-based wealth predictions in downstream tasks. **e** Cross-sectional relationship between average maximum temperature and wealth across survey locations, as estimated with survey wealth data (black) and from satellite-based models. Each line is a bootstrap of the cross-sectional regression (100 bootstraps, sampling villages with replacement). Best-performing models recover temperature-wealth relationships that are closest to estimates using ground-measured data, and CNN-based models perform much better than scalar nightlights models. **f** Evaluation of hypothetical targeting program in which all villages below some desired threshold in the asset distribution receive the program and villages above the threshold do not.

have not also been additionally jittered. This suggests that the true (unobserved) performance of our main results is higher than we report, given that we are evaluating on data that have been jittered. Extrapolating backward to a hypothetical setting of no jitter in training data suggests that locational noise in ground data is reducing model performance by $r^2 = 0.07$, roughly the difference between our best and worst performing CNN models (Fig. 1).

Downstream tasks To demonstrate the applicability of our satellite-based estimates to downstream research or policy tasks, we consider two use cases. The first is understanding why some locations are wealthier than others. Here we study associations between wealth and exposure to extreme temperatures, as much past work has indicated the wealth-temperature relationship is nonlinear [7, 25], and because temperature data are readily available for all study locations in an independent gridded dataset [12]. Ground-based survey data indicate a non-linear relationship between village-level wealth and maximum temperature in the warmest month, and out-of-country estimates from CNN-based models recover this relationship very closely (Fig. 2e); estimates from simple scalar nightlights models do not. While none of these cross-sectional estimates are well suited for causal identification of the impact of temperature on wealth [7, 8], we view the close match between satellite- and ground-based estimates of the temperature-wealth relationship as evidence that satellite-based estimates can be useful for these types of research questions.

We also use our estimates to evaluate the hypothetical targeting of a social protection program (e.g., a cash transfer), in which all villages below some asset level receive the program and villages above the threshold do not. Targeting on survey-derived asset data is a common program disbursement approach in developing countries [17]. We compare targeting accuracy, defined as the percent of villages receiving the correct program, using estimates from different satellite-based models, under the assumption that survey-based ground data describe the "true" asset distribution. Our models again perform well on this task. For instance, using MS+NL estimates to allocate a program to households below median wealth yields a targeting accuracy of 81%, versus 75% for a CNN Transfer model and 62% for a scalar nightlights model (Fig. 2f). Importantly, these estimates likely understate "true" targeting accuracy given that ground data are themselves measured with some noise.

Scalability To demonstrate the scalability of our overall approach, we construct a 7.65km/pixel gridded wealth map of Nigeria, Africa's most populous country, for the years 2012-2014 using our CNN MS+NL model (Fig. 3). Visualizing both inputs and model predictions shows how our model

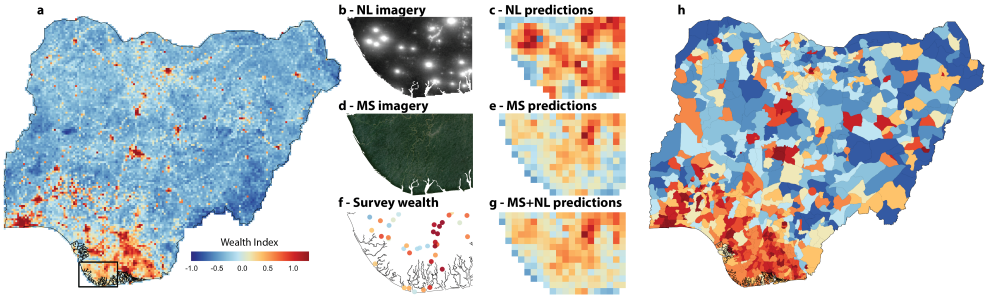

Figure 3: **Spatial extent of imagery allows wealth predictions at scale**. **a** Satellite-based wealth estimates across Nigeria at pixel level. **b, d** Imagery inputs to model. **f** Ground truth input to model. **c, e, g** Model predictions with just nightlights (NL) as input, just multispectral (MS) imagery as input, and the concatenated NL and MS features as input. In this region, the model appears to rely more heavily on MS than NL inputs, ignoring light blooms from gas flares visible in (b). **h** Deciles of satellite-based wealth index across Nigeria, population weighted using Global Human Settlement Layer population raster, and aggregated to Local Government Area level.

learns to combine the two inputs, for example ignoring very bright nightlights pixels associated with oil flaring in the southern part of the country that are not also associated with high wealth (Fig. 3b-g). Pixels are easily aggregated to higher administrative units using existing population rasters, and show strong latitudinal gradients of wealth across the country (Fig. 3h).

Generating the pixel-level raster involves processing $\sim$9.1 billion pixels of daytime and nighttime imagery and takes $<$30hrs from raw imagery inputs, including 4hrs of model training on a NVIDIA Titan X GPU (excluding hyperparameter search) and roughly 24hrs for imagery processing and raster generation. By comparison, a nationally representative household survey typically takes months to years to execute, at an average cost of $1-2million USD [11]. While this comparison does not imply that our approach can replace household surveys, our approach can speed up estimation of local-level wealth where survey data are unavailable.

6 Conclusion

Our satellite-based deep learning approach to measuring asset wealth is both accurate and scalable, and consistent performance on held-out countries suggests that it could be used to generate wealth estimates in countries where data are unavailable. Results suggest that such estimates could be used to help target social programs in data poor environments, as well as to understand the determinants of variation in well-being across the developing world.

However, while our CNN-based approach outperforms approaches to poverty prediction that use simpler features common in the literature (e.g., scalar nightlights [20]), the CNN model is less interpretable than these simpler approaches, perhaps inhibiting adoption by the policy community. A key avenue for future research is in improving the interpretability of deep learning models in this context, and in developing approaches to navigate this apparent performance-interpretability tradeoff.

Our deep learning approach is also perhaps best viewed as a way to amplify rather than replace ground-based survey efforts, as local training data can often further improve model performance (Fig. 1b), and because other key livelihood outcomes often measured in surveys—such as wealth distribution within households, or between households within villages—are more difficult to observe in imagery. Similarly, our approach could also be applied other key outcomes, including consumption-based poverty metrics or health indicators. Performance in these related domains will depend both on the availability and quality of training data, which remains limited for key outcomes such as consumption in most geographies. Finally, our approach could likely be further improved by the incorporation of higher-resolution optical and radar imagery now becoming available at near daily frequency, or in combination with data from other passive sensors such as mobile phones [5] or social media platforms [30]. All represent scalable opportunities to expand the accuracy and timeliness of data on key economic indicators in the developing world, and could accelerate progress towards measuring and achieving global development goals.

Broader Impact

Accurate, local-level measurement of economic livelihoods is key to understanding and affecting economic development. We show how wealth in African villages across dozens of countries can be accurately predicted by a machine learning model using publicly-available satellite imagery as input. We demonstrate how our approach is useful in explaining both spatial differences and temporal changes in livelihoods, and how it is scalable across broad geographies. We also show how derived estimates can be used in research and policy applications.

We expect that methods from this work may help governmental and non-governmental organizations provide more targeted and data-driven distributions of resources (e.g., cash transfers) to villages and districts with the greatest need, as we demonstrated in our paper. However, we also recognize that our models are imperfect; for example, should our models actually be deployed in practice, regions where our model makes mistakes may be put at a disadvantage. Thus, decisions should not be made based on the outputs of our models alone. Our models are only as good as the data they are trained on and therefore reflect any sampling and/or survey biases that may exist in the underlying survey data.

Relevance to NeurIPS ML for Economic Policy Workshop

Our work provides a clear demonstration of the utility of machine learning models (specifically convolutional neural networks) for estimating local-level economic well-being, which is critical data for making effective economic policy. We show how our models can shed light on under-developed regions in sub-Saharan Africa, especially in regions where traditional survey data are extremely scarce or nonexistent. We have also released our data and code publicly for everyone to engage and test their own models with.

Note: a previous version of this paper was published earlier in 2020 [33]. However, a member of the workshop organizing committee gave explicit permission for this paper to be submitted to the 2020 NeurIPS ML for Economic Policy Workshop.

References

[1] Klaus Ackermann et al. "Object Recognition for Economic Development from Daytime Satellite Imagery". In: *arXiv preprint arXiv:2009.05455* (2020).

[2] Sabina Alkire et al. *Multidimensional poverty measurement and analysis*. Oxford University Press, USA, 2015.

[3] G Azzari and D B Lobell. "Landsat-based classification in the cloud: An opportunity for a paradigm shift in land cover monitoring". In: *Remote Sensing of Environment* (May 2017), pp. 1–11.

[4] Boris Babenko et al. "Poverty Mapping Using Convolutional Neural Networks Trained on High and Medium Resolution Satellite Images, With an Application in Mexico". In: *NIPS 2017 Workshop on Machine Learning for the Developing World*. 2017. URL: https://arxiv.org/abs/1711.06323.

[5] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. "Predicting poverty and wealth from mobile phone metadata". In: *Science* 350.6264 (2015), pp. 1073–1076.

[6] Marshall Burke, Sam Heft-Neal, and Eran Bendavid. "Sources of variation in under-5 mortality across sub-Saharan Africa: a spatial analysis". In: *The Lancet Global Health* 4.12 (2016), e936–e945.

[7] Marshall Burke, Solomon M Hsiang, and Edward Miguel. "Global non-linear effect of temperature on economic production". In: *Nature* 527.7577 (2015), p. 235.

[8] Melissa Dell, Benjamin F Jones, and Benjamin A Olken. "What do we learn from the weather? The new climate-economy literature". In: *Journal of Economic Literature* 52.3 (2014), pp. 740–98.

[9] Christopher D Elvidge et al. "VIIRS night-time lights". In: *International Journal of Remote Sensing* 38.21 (Nov. 2017), pp. 5860–5879. ISSN: 13665901. DOI: 10.1080/01431161.2017.1342050. URL: https://www.tandfonline.com/doi/full/10.1080/01431161.2017.1342050.

[10] Ryan Engstrom, Jonathan Samuel Hersh, and David Locke Newhouse. "Poverty from space: using high-resolution satellite imagery for estimating economic well-being". In: 1 (Dec. 2017), pp. 1–36. URL: http://documents.worldbank.org/curated/en/610771513691888412/Poverty-from-space-using-high-resolution-satellite-imagery-for-estimating-economic-well-being.

[11] Jessica Espey et al. "Data for development: a needs assessment for SDG monitoring and statistical capacity development". In: *Sustainable Development Solutions Network. Available at http://unsdsnorg/wp-content/uploads/2015/04/Data-for-Development-Full-Report.pdf* (2015).

[12] Stephen E Fick and Robert J Hijmans. "WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas". In: *International Journal of Climatology* 37.12 (2017), pp. 4302–4315.

[13] Deon Filmer and Lant H Pritchett. "Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India". In: *Demography* 38.1 (2001), pp. 115–132.

[14] Deon Filmer and Kinnon Scott. "Assessing Asset Indices". In: *Demography* 49 (2012), pp. 359–392.

[15] Noel Gorelick et al. "Google Earth Engine: Planetary-scale geospatial analysis for everyone". In: *Remote Sensing of Environment* (2017). DOI: 10.1016/j.rse.2017.06.031. URL: https://doi.org/10.1016/j.rse.2017.06.031.

[16] Nicholas Graetz et al. "Mapping local variation in educational attainment across Africa". In: *Nature* 555.7694 (2018), p. 48.

[17] Margaret E Grosh et al. *For protection and promotion: The design and implementation of effective safety nets*. The World Bank, 2008.

[18] Kaiming He et al. "Identity Mappings in Deep Residual Networks". In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Springer International Publishing, Oct. 2016, pp. 630–645. DOI: 10.1007/978-3-319-46493-0_38. URL: http://link.springer.com/10.1007/978-3-319-46493-0%7B%5C_%7D38%20https://arxiv.org/abs/1603.05027.

[19] Andrew Head et al. "Can Human Development Be Measured with Satellite Imagery?" In: *Proceedings of the Ninth International Conference on Information and Communication Technologies and Development*. ICTD '17. Lahore, Pakistan: ACM, 2017, 8:1–8:11. ISBN: 978-1-4503-5277-2. DOI: 10.1145/3136560.3136576. URL: https://dl.acm.org/citation.cfm?doid=3136560.3136576.

[20] J Vernon Henderson, Adam Storeygard, and David N Weil. "Measuring economic growth from outer space". In: *American economic review* 102.2 (2012), pp. 994–1028.

[21] Feng-Chi Hsu et al. "DMSP-OLS Radiance Calibrated Nighttime Lights Time Series with Intercalibration". In: *Remote Sensing* 7.2 (Feb. 2015), pp. 1855–1876. DOI: 10.3390/rs70201855. URL: http://www.mdpi.com/2072-4292/7/2/1855.

[22] Wenjie Hu et al. "Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. Honolulu, HI, USA: Association for Computing Machinery, 2019, pp. 353–359. ISBN: 9781450363242. DOI: 10.1145/3306618.3314263. URL: https://doi.org/10.1145/3306618.3314263.

[23] Neal Jean et al. "Combining satellite imagery and machine learning to predict poverty". In: *Science* 353.6301 (2016), pp. 790–794. ISSN: 0036-8075. DOI: 10.1126/science.aaf7894.

[24] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[25] William D Nordhaus. "Geography and macroeconomics: New data and new findings". In: *Proceedings of the National Academy of Sciences* 103.10 (2006), pp. 3510–3517.

[26] Aaron Osgood-Zimmerman et al. "Mapping child growth failure in Africa between 2000 and 2015". In: *Nature* 555.7694 (2018), p. 41.

[27] Neeti Pokhriyal and Damien Christophe Jacques. "Combining disparate data sources for improved poverty prediction and mapping". In: *Proceedings of the National Academy of Sciences* 114.46 (2017), E9783–E9792.

[28] Robert C Reiner Jr et al. "Variation in Childhood Diarrheal Morbidity and Mortality in Africa, 2000–2015". In: *New England Journal of Medicine* 379.12 (2018), pp. 1128–1138.

[29] David E Sahn and David Stifel. "Exploring alternative measures of welfare in the absence of expenditure data". In: *Review of income and wealth* 49.4 (2003), pp. 463–489.

[30] Evan Sheehan et al. "Predicting economic development using geolocated Wikipedia articles". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2698–2706.

[31] Lucy S Tusting et al. "Mapping changes in housing in sub-Saharan Africa from 2000 to 2015". In: *Nature* (2019), p. 1.

[32] Gary R. Watmough et al. "Socioecologically informed use of remote sensing data to predict rural household poverty". In: *Proceedings of the National Academy of Sciences* 116.4 (Jan. 2019), pp. 1213–1218. ISSN: 0027-8424. DOI: 10.1073/pnas.1812969116. URL: https://www.pnas.org/content/116/4/1213.

[33] Christopher Yeh et al. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa". In: *Nature Communications* 11.1 (May 2020), p. 2583. ISSN: 2041-1723. DOI: 10.1038/s41467-020-16185-w. URL: https://doi.org/10.1038/s41467-020-16185-w.